УДК 17.024 004.8 + 167/168 + 159.9 + 316.6

К ВОЗМОЖНОСТИ МОРАЛИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

С. И. Голенков

Самара, Самарский университет

Аннотация. Вопрос о возможности морали искусственного интеллекта (ИИ) имеет две стороны. Первая — мораль для ИИ, создаваемая разработчиками ИИ, теоретиками и практиками права и этики. Вторая — создается самим ИИ для себя. Обосновывается принципиальная невозможность создания морали самим ИИ для себя. Мораль как способ регуляции человеческого поведения в своем основании опирается на разум человека, его интеллектуальные способности. Интеллект человека может действовать в двух режимах — «творческом» и «вычислительном». «Творческий» производит сущее из Небытия свободно, «вычислительный» создает сущее из уже существующих сущих или их фрагментов путем совершенствования и/или комбинирования по алгоритму. Искусственный интеллект может действовать только по алгоритму, потому мораль для него всегда будет иметь человеческие основания.

Ключевые слова: мораль, искусственный интеллект, естественный интеллект, интеллект-событие, интеллект-вычисление

TO THE POSSIBILITY OF MORALS OF ARTIFICIAL INTELLECT

Sergey Golenkov

Samara, Samara University

Abstract. The question of the possibility of morality of artificial intelligence (AI) has two sides. The first is morality for AI, created by AI developers, theorists and practitioners of law and ethics. The second is created by the AI itself for itself. The fundamental impossibility of creating morality by AI for itself is substantiated. Morality as a way of regulating human behavior is based

on the human mind, his intellectual abilities. Human intelligence can operate in two modes — “creative” and “computational”. The “creative” produces essence from Nothingness freely, the “computational” constructs essence from already existing beings or their fragments by means of improvement and / or combination according to the algorithm. AI can only act according to an algorithm, because morality for it will always have human grounds.

Key words: morality, artificial intelligence, natural intelligence, intelligence-event, intelligence-computing

Вопрос о морали искусственного интеллекта (ИИ) возник в первой половине прошлого века, и, еще до своего теоретического обоснования, его решение было предложено Айзеком Азимовым в законах робототехники. С тех пор дискуссии о ее возможности перешли из плоскости обсуждений в чисто практическую плоскость. Возрастающая актуальность этической повестки в области ИИ обусловлена как достижениями робототехники, так и растущим вниманием к этим достижениям со стороны государств и общества.

Вопрос о морали ИИ имеет две стороны. Во-первых, речь может идти о морали *для* ИИ. В русле именно такого понимания разработчики создают архитектуру ИИ исходя из *человеческих* представлений о допустимом и недопустимом, норме и девиации, добре и зле. Вторая сторона вопроса высвечивается как возможность ИИ создавать мораль самому себе. В этом случае возникает вопрос: может ли появиться мораль у ИИ, отличная от морали человеческой в том смысле, что ИИ *сам для себя* становится источником морали?

Начиная с Сократа и до сегодняшнего дня вопрос о морали всегда рассматривался в связи с вопросом о разумности, так как считалось и до сих пор считается, что мораль возможна лишь у разумных существ. В таком ракурсе возможность морали у ИИ напрямую зависит от разумности ИИ. Является ли разумность ИИ тождественной разумности интеллекта человеческого, то есть естественного (ЕИ)? Сопоставление этих интеллектов покоится на молчаливом признании тезиса о том, что ИИ и ЕИ являются лишь *видами одного и того же* интеллекта.

Вокруг темы о разумности ИИ ведутся жаркие споры, которые поделили спорящих на оптимистов и скептиков. Первые считают

ИИ не просто тождественным ЕИ: в ближайшее время он сравняется с ним в интеллектуальных способностях, а уже в обозримой перспективе, в точке «технологической сингулярности» (В. Вернон), превзойдет человеческий аналог. Истоки такого оптимизма заложены Аланом Тьюрингом, который еще в середине прошлого столетия позитивно оценивал возможности машинного интеллекта и считал, что они не уступают интеллекту человека. Его уверенность покоилась на убеждении в существовании глубоких связей между деятельностью нервной системы человека и работой цифровых вычислительных машин [4, с. 17]. Свой оптимизм сторонники сильного ИИ основывают на поразительных достижениях последних 20 лет в сферах творческой деятельности человека, до сих пор считавшихся исключительно его прерогативой.

Оптимисты признают наличие проблем и рисков в области морали, связанных с созданием сильного ИИ. Однако они считают, что проблемы решаемы, а риски контролируемы. Оптимисты активно обсуждают типы перехода от человеческого интеллекта к сильному ИИ. Сторонники преемственности в развитии интеллекта полагают, что разрыв между ЕИ и ИИ будет непринципиальным. Джозеф Кораби пишет о том, что ИИ в использовании своих возможностей будет ограничен в областях «скептических проблем», то есть в ситуациях, когда принципиально невозможно сделать рациональный выбор между возможностями, основываясь только на интеллектуальных способностях разума. В такой ситуации, считает Кораби, ИИ может быть парализован «из-за проблем и конфликтов в своих собственных мотивационных схемах» [5, с. 5].

В начале нынешнего столетия возникла и начала оформляться концепция *дружественного искусственного интеллекта* (ДИИ). Философия ДИИ покоится на предпосылке, что носители ДИИ не только не будут приносить вреда человечеству, но и будут всемерно оказывать материально-информационную поддержку людям, вплоть до полного обеспечения желаний и потребностей каждого отдельно взятого человека. Бен Гёрцель считает, что сосуществование ЕИ и ИИ будет способствовать формированию морали у ИИ и, кроме того, повлечет за собой изменение морали человека [6].

Скептики, в свою очередь, ссылаются на принципиальное различие между ЕИ и ИИ. Они указывают на два важных обстоятельства, не позволяющих считать достоверной методологию оценки уровня развития ИИ машины, предложенную Тьюрингом. Во-первых, вопрос об имитации мышления машиной смещает внимание с самого ИИ на его результаты. Во-вторых, само мышление оптимистами считается вычислительным без достаточных оснований. Джон Сёрл представил мысленный эксперимент, известный как аргумент «Китайская комната» [7]. Анализируя его, Сёрл дает отрицательный ответ на вопрос о возможности аналогии ИИ с ЕИ. Он считает, что компьютерные программы ИИ реализуются как алгоритмические операции вычисления по формально определенным правилам и элементам, тогда как ментальные операции понимания являются проявлением интенциональности особого биологического субстрата — человеческого мозга. Джозеф Вейценбаум настаивает, что даже если машина имитирует человеческую деятельность, например работает с информацией, то эта деятельность у человека и вычислительной машины протекает принципиально по-разному [1, с. 144–145].

Отличия ИИ от ЕИ становятся явными в творческой деятельности. Исследователи различают три вида творчества: доведение чего-либо до *совершенства*, *рекомбинация* уже существующего и творение «*словно ниоткуда*» (Маркус дю Сотой) [3]. Первые два вида производства нового к настоящему времени не составляют труда для ИИ. Машины успешно обыгрывают человека в настольные игры, пишут стихи и портреты, ваяют скульптуры и пишут музыку. В основании видов совершенствования и рекомбинации у ИИ лежат большие базы данных (текстовых, визуальных, музыкальных, живописных и т. д.), «самообучение» и алгоритм их анализа и построения. Оба этих вида творчества развиваются по пути наращивания вычислительных мощностей ИИ, создавая новое из *уже существующего*.

Творение «ниоткуда» принципиально отличается от творения по видам совершенствования и рекомбинации. В производстве «ниоткуда» вычисление не участвует, потому что, пока не возникло новое, невозможно использование алгоритма, составляющего суть вычисления. Следуя тысячелетней традиции, современные

исследователи источником творчества считают Небытие. Анализируя творческие моменты человеческого существования, В. А. Конев пишет: «Творчество — это бытийный акт, акт творения бытия. С онтологической точки зрения творчество есть прорыв небытия в бытие. Небытие (бездна) есть причина, исток творчества. Узреть небытие (ничто как нечто) и заполнить его своим произведением — вот онтологическое назначение творчества» [2, с. 152].

Согласно В. А. Коневу принципиальное отличие творения нового «из Небытия» от производства нового по пути совершенствования и рекомбинации заключается в том, что творение из «*уже* существующего» есть завершение прошлого, тогда как творение из «*еще не* существующего» вводит будущее. Кроме того, способы творения из «*уже* есть» и из «*еще не* есть» различаются в своей реализации методически. Первый укладывается в правило «основание–следствие», которое ведет от уже существующего к новому существующему [Там же, с. 154, 156]. Второй же, строго говоря, *не имеет правила* и реализуется по методу «вдруг» [Там же, с. 153–154], то есть имеет природу события.

Различие модусов творчества вычисления (совершенствование, рекомбинация) и события (из Небытия «вдруг») производят соответственно различие модусов интеллекта — вычисления и события. Если воспользоваться этим различением, то можно говорить, что ЕИ человека включает в себя оба способа интеллекта — вычислительное и событийное, тогда как ИИ используют только интеллект-вычисление.

Модальности интеллекта по-разному участвуют в производстве мысли. Интеллект-событие «запускает» сущее, собирая его из гетерогенного «материала». Дело же интеллекта-вычисления состоит в «работе» с уже возникшим сущим, в его развитии, совершенствовании и применении. Иначе говоря, интеллект-событие выступает началом интеллекта-вычисления. В понимании принципов «работы» интеллекта-вычисления многое сделано герменевтикой и создателями ИИ.

В итоге можно определенно говорить, что мораль человека (ЕИ) формируется обоими модусами интеллекта, тогда как мораль ИИ может производиться только интеллектом-вычислением. Можно

утверждать, что мораль ИИ — это всегда мораль *для* ИИ, так как своим источником может иметь лишь мораль человеческую, сам ИИ не может быть источником собственной морали.

ЛИТЕРАТУРА

1. *Вейценбаум Дж.* Возможности вычислительных машин и человеческий разум: От суждений к вычислениям. М. : Радио и связь, 1982.

2. *Конев В. А.* Тайна «Да будет», или «…Круговая черта по лицу бездны» // Конев В. А. Смыслы культуры : сб. ст. Самара : Изд-во «Самар. ун-т», 2016.

3. Математик из Оксфорда объясняет, как ИИ может повысить творческие способности человека. Искусство и инновации в эпоху искусственного интеллекта. URL: https://www.theverge.com/2019/4/10/18303438/artificial-intelligence-ai-art-book-interview-marcus-du-sautoy-the-creativity-code (дата обращения: 13.03.2019).

4. *Тьюринг А.* Может ли машина мыслить? (С приложением статьи Дж. фон Неймана «Общая и логическая теория автоматов»). М. : ГИФМЛ, 1960.

5. *Corabi J.* Superintelligent AI and Skepticism // Journal of Evolution and Technology. 2017. Vol. 27, iss. 1. June. P. 4–23.

6. *Goertzel B.* Superintelligence: Fears, Promises and Potentials: Reflections on Bostrom's Superintelligence, Yudkowsky's From AI to Zombies, and Weaver and Veitas's "Open-Ended Intelligence" // Ibid. 2015. Vol. 24, iss. 2. Nov. P. 55–87.

7. *Searle J. R.* Minds, brains, and programs // Behavioral and Brain Sciences. 1980. 3 (3). P. 417–457. URL: http://cogprints.org/7150/1/10.1.1.83.5248.pdf (дата обращения: 25.01.2018).